ANALYSIS OF COUNTS WITH TWO LATENT CLASSES, WITH APPLICATION TO RISK ASSESSMENT USING PHYSICIAN VISIT RECORDS

by

Huijing Wang

B.Sc., Simon Fraser University, 2010

A Thesis submitted in partial fulfillment
of the requirements for the degree of

Master of Science

in the
Department of Statistics and Actuarial Science
Faculty of Science

SIMON FRASER UNIVERSITY
Summer 2012

APPROVAL

Name: Huijing Wang

Degree: Master of Science

Title of Thesis: ANALYSIS OF COUNTS WITH TWO LATENT CLASSES, WITH APPLICATION TO RISK ASSESSMENT USING PHYSICIAN VISIT RECORDS

Examining Committee:

Dr. Tim B. Swartz
Professor
Chair

Dr. X. Joan Hu
Professor
Senior Supervisor

Dr. Brad McNeney
Associate Professor
Supervisor

Dr. John J. Spinelli
Adjunt Professor
External Examiner

Date Approved: _____

Abstract

Motivated by the CAYACS program at BC Cancer Research Center, this thesis project introduces a latent class model to formulate event counts. In particular, we consider a population with two latent classes, such as an at-risk group and a not-at-risk group of cancer survivors in the CAYACS program. Likelihood-based inference procedures are proposed for estimating the model parameters with or without one class fully specified. The EM algorithm is adapted to compute the MLE; a pseudo MLE of the model parameters is proposed to reduce computing intensity and improve inference efficiency using readily available supplementary information. The estimation procedures are studied via simulation regarding both efficiency and robustness. We illustrate the methodology with the physician claim data of the CAYACS cohort for risk assessment throughout the project. With the latent class model, we identify risk factors for cancer survivors to late and on-going problems and obtain an alternative, perhaps more desirable, comparison of the cohort with the general population.

Acknowledgments

First and foremost I would like to express my sincerest gratitude to my supervisor, Dr. Joan Hu. I took my very first statistics course at SFU with her, which inspired and encouraged me to go on the path of statistics. Her excellent guidance, caring nature, patience, knowledge and inspiration has supported me both academically and mentally throughout my two year masters study. I attribute the level of my masters degree to her encouragement and effort and without her this thesis, too, would not have been completed or written. But most importantly, she believed in it before I had believed myself that I could do this.

I would never have thought that I could find an area to be passionate about that would change my life so much before I learned statistics. I would like to thank the whole Department of Actuarial Science and Statistics, always feeling fortunate and thankful that I can be with such an intelligent group. Thank-you to the departmental professors who have provided me professional training in statistics. I enjoyed every single lecture and seminar attended here. Our department has provided me with an excellent atmosphere for doing research. I give my best appreciation to my committee, Drs. Brad McNeney and John Spinelli. Thank you for your wonderful advice and valuable time. Many thanks to all my graduate fellows whose help and friendship make my study life delightful, especially to Lilian Xia, Rose Yu, Jean Shin, Jie Liu, Zheng Sun, Hua Zheng, Jinny Lim, Yingying Chen, Joslin Goh, Shirin Golchi, Oksana Chrebtii, Andrew Henrey, Fabian Moya and Harsha Perera. I will never forget all the great moments we shared and spent, and how understanding, encouraging, humorous and kind you all are.

I am very grateful for all the personal and technical helps from the CAYACS group at the BC Cancer Agency. Particularly, I thank the PI of the CAYACS program, Mary McBride, for her kindness and generous support. The BC Cancer Agency, the BC Childrens Hospital and the BC Ministry of Health approved access to and use of the data facilitated

by Population Data BC for this study.

I feel deeply indebted to my parents. Despite leaving home for many years, they have never complained or stopped giving their everlasting love, care and support to me unconditionally. I hope my effort can let them be proud.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background and Motivation

The population of cancer survivors has been increasing rapidly due to improvements in cancer treatments. These survivors are often at risk for late and ongoing problems that are mainly treatment-related. To determine care needs and required resources and to evaluate/develop strategies for long-term management, there have been demands of risk assessment particularly for those diagnosed with cancer at a young age. The Childhood, Adolescent, Young Adult Cancer Survivor (CAYACS) research program at British Columbia Cancer Research Center (*http://www.cayacs.ca*), using existing population-based datasets and record linkage methodology, has been conducting a series of epidemiologic, clinical and health service studies relating to survivorship issues of cancer survivors diagnosed at age 0 to 19; see McBride *et al.* (2010).

One of CAYACS's specific objectives is to evaluate the physician visit frequency and cost of young cancer survivors in British Columbia and to identify the risk factors, in contrast to the general population. McBride *et al.* (2011) report an analysis of the available physician claim data associated with a cohort of young cancer survivors, compared to the general population, to identify the risk factors for the physician visits and cost. They find that demand for physician care among the young cancer survivors is considerably greater than it from the general population within similar age and sex group. The analysis provides insights into physician visit patterns of the survivors and, at the same time, raises issues to be further explored. For example, the comparison of all the cancer survivors to the general population may implicitly reveal whether the portion of survivors in the cohort at

risk for late and ongoing problems is larger than that in the general population. It does not appropriately assess the survivor cohort's risk rate as the consequences of the original cancer diagnoses. For another example, the analysis in McBride *et al.* (2011) indicates that females have significantly higher physician visit frequency than males in the cohort. It is not clear whether this identifies sex as a risk factor or simply reflects a pattern of physician visits over all, since this pattern is also seen in the general population.

Some preliminary analysis indicates that, while many survivors visit physicians rather frequently, some survivors in the cohort show a similar physician visit frequency to people from the general population. This leads one to conjecture that the cohort is a mixture of "at-risk" and "not-at-risk" groups. The individuals in the "at-risk" group are those who are suffering the later effects of the original cancer diagnosis and with potentially higher rate of physician visits, while individuals in the "not-at-risk" group are not at an increased risk and have the same physician visit pattern as the general population. Evaluating the at-risk group in the cohort separately may provide a better assessment of the risk to late and ongoing problems of cancer survivors; the risk factors can be then identified via the associated regression analysis. However, the membership of a subject in either the at-risk group or the not-at-risk group is not observable. This motivated us to consider a latent class model with two latent classes: at-risk and not-at-risk groups.

Goodman (1974) formalizes the latent class modeling introduced by Lazarsfeld and Henry (1968), and derives the maximum likelihood estimation procedure. Latent class modeling has had a wide range of applications; see, for example, Magidson and Vermunt (2002); Pepe and Janes (2007); Vermunt (2008). The formulation gives a convenient framework for risk assessment, to study the features of physician visit patterns particularly due to the late and ongoing treatment-related problems of cancer survivors. It also leads to a natural comparison of the survivors in the at-risk group to the general population. We assume the "not-at-risk" group in the latent class modeling has exactly the same distribution of physician visits as the general population.

In the analysis with a latent class model, one needs to specify the underlying probability model into a parametric form for each of the latent classes to avoid a non-identifiability problem in general. On the other hand, in addition to other issues such as computational robustness when implementing the likelihood-based procedures with latent class models (e.g., Hall and Shen, 2010), the efficiency of the MLE will drop considerably due to the increasing number of parameters. A model with two latent classes has almost 3 times as

many parameters as a single marginal model. To address this issue, we make efforts to utilize the potentially available rich information on the general population in this project.

We firstly assume that the distribution of the not-at-risk group is fully specified. The estimated distribution from the general population is used as the true distribution in the real data analysis. This allows us to focus on making inferences about the at-risk group, to develop a more efficient inferential procedure with a partially specified model for the at-risk group. At the same time, it reduces the computational intensity of the likelihood-based approach. We then explore how to account for the variation of the estimated distribution for the class of the not-at-risk cancer survivors.

This thesis project aims at formulating the CAYACS physician visits by a latent class model, and developing the associated likelihood-based estimating procedures, to assess the frequency of physician visits of the group at-risk and identify the associated risk factors. We employ the CAYACS program to motivate and illustrate the proposed modeling and inferential procedures. However, the methodology is not limited to the program and can be applied rather broadly.

1.2 Framework and Model Specification

Let N and Z be a subject's count of physician visits over the time period $(0, T]$ and covariate vector, respectively. Here the observation period in the particular CAYACS application is the time interval starting from when a BC resident diagnosed with cancer at a young age becomes a "survivor" until his/her death or the end of the data collection. To formulate the cohort's potential two strata, we consider a cohort with two latent classes, corresponding to the "at-risk" and "not-at-risk" groups in the cancer survivor cohorts. Introduce a latent binary variable η to indicate if a subject belongs to the group at-risk. Assume that $E(\eta|Z) = P(\eta = 1|Z)$ is $p(Z; \alpha)$, known up to the parameter vector α. This allows us to identify risk factors with the group at-risk, in which subjects have a pattern of physician visits different from the general population. It leads to a finite mixture modeling as follows.

We specify the underlying probability models for N with the groups at-risk and not-at-risk as Poisson distributed with means $E(N|\eta = 1, T, Z) = \Lambda_1(T, Z; \beta)$ and $E(N|\eta = 0, T, Z) = \Lambda_0(T, Z; \theta)$, respectively. The popular zero-inflated Poisson (ZIP) model (e.g., Lambert, 1992; Hall and Shen, 2010) is a special case. Suppose that the cohort of interest has n independent subjects. We allow the observation period to vary from subject to

subject. Denote the time lengths, observed counts, at-risk indicators and covariates associated with the subjects by $\{(T_i, N_i, \eta_i, Z_i) : i = 1, \ldots, n\}$. Here N_i is the observed count of subject i's physician visits over time interval $(0, T_i]$. Our primary interest is in estimating the parameters α, β and θ in $p(Z;\alpha)$, $\Lambda_1(T, Z;\beta)$ and $\Lambda_0(T, Z;\theta)$ with the available data $\{(T_i, N_i, Z_i) : i = 1, \ldots, n\}$. For example, an estimator of α can be used to evaluate $p(Z;\alpha)$, which gives a numerical risk assessment, or to identify a risk factor by conducting a significance test on the effect of a covariate. An estimator of β, on the other hand, can be used to identify factors associated with high visit frequency in the group at-risk. Commonly used parametric specifications are the ones used in the logistic and loglinear models: $logit\{p(Z;\alpha)\} = \alpha' Z$, $log\{\Lambda_1(T, Z;\beta)\} = \beta_0 + \beta_1' Z + \beta_2 T$ and $log\{\Lambda_0(T, Z;\theta)\} = \theta_0 + \theta_1' Z + \theta_2 T$.

In the CAYACS application, the available physician claim data are directly about the conditional distribution of N, given T and Z, a mixture Poisson distribution:

$$P(N|T, Z) = P(N|\eta = 1, T, Z;\beta)p(Z;\alpha) + P(N|\eta = 0, T, Z)[1 - p(Z;\alpha)].$$

This may lead to some rather intensive computing in the statistical analysis. There exists usually a large amount of information on the general population. Taking the distribution of the not-at-risk group as the same as that of the general population, we start with assuming that $P(N|\eta = 0, T, Z)$, the distribution of N conditional on $\eta = 0$ and (T, Z), is known, and focus on making inferences about (α, β). In the actual data analysis, we take the estimate of the distribution obtained using the information from the general population as the true value. We then explore how to account for the variation of the estimator.

1.3 Project Outline

The rest of this thesis is organized as follows. Chapter 2 presents inference procedures assuming the distribution of physician visit counts in the "not-at-risk" group is known due to the readily available rich information from the general population. We present three sets of maximum likelihood estimators obtained by maximizing full-data likelihood, maximizing observed-data likelihood, and an application of EM-algorithm. The first set is not practically useful. It is used to study the properties of the last two sets, which are theoretically equivalent.

The discussion at the end of Chapter 2 motivates Chapter 3 to propose estimation procedures with both latent classes not-fully specified. We start with making inferences for all parameters (α, β, θ) by the maximum likelihood estimation using the survivor cohort data. Then we consider a consistent estimator of θ from the general population and plug it in the likelihood functions to attain a pseudo MLE of (α, β). We take account for the variation of the estimation from the general population for the distribution of the not-at-risk group in estimating the standard deviation of the pseudo MLE.

Both chapters describe and report the simulation studies conducted to examine the finite sample properties. To illustrate the methodology, real CAYACS physician visit data are analyzed in Chapter 4 via the approaches proposed in the previous chapters. Final remarks are given in Chapter 5.

Chapter 2

Inference Procedures with One Class Fully Specified

In this chapter, we assume the distribution of the not-at-risk group is known and derive likelihood-based approaches for estimating the parameters (α, β) in the aforementioned model for the group at-risk. An application of the EM algorithm is presented in the chapter. A simulation study was conducted to investigate efficiency and robustness of the estimation procedures. We show at the end of this chapter that the inferences on (α, β) can be biased and inefficient if an estimate from the general population is used as the distribution of the not-at-risk group. This motivates the discussion in Chapter 3.

2.1 Likelihood Functions and EM Algorithm

Consider the mixture Poisson model introduced in *Section 1.2*: conditional on η and (T, Z), $N \sim Poisson(\Lambda_\eta(T, Z))$. We assume here that $\Lambda_0(T, Z)$ is known and $\Lambda_1(T, Z)$ is specified as $\Lambda_1(T, Z; \beta)$ up to parameter β. This model reduces to the zero-inflated Poisson (ZIP) model (e.g., Lambert, 1992) when $\Lambda_0(T, Z) \equiv 0$.

Under the mixture Poisson model, the likelihood function of (α, β) based on the available data $\big\{(T_i, N_i, Z_i) : i = 1, \ldots, n\big\}$ is

$$\begin{aligned} L(\alpha, \beta; N|T, Z) &= \prod_{i=1}^{n} \Big\{ P(N_i \big| \eta_i = 1, T_i, Z_i; \beta) P(\eta_i = 1 | Z_i; \alpha) \\ &\quad + P(N_i \big| \eta_i = 0, T_i, Z_i)[1 - P(\eta_i = 1 | Z_i; \alpha)] \Big\}, \end{aligned} \tag{2.1}$$

where $P(N_i|\eta_i = 0, T_i, Z_i) = \Lambda_0(T_i, Z_i)^{N_i} e^{-\Lambda_0(T_i,Z_i)}/N_i!$ and $P(N_i|\eta_i = 1, T_i, Z_i; \beta) = \Lambda_1(T_i, Z_i; \beta)^{N_i} e^{-\Lambda_1(T_i,Z_i;\beta)}/N_i!$. The maximum likelihood estimator (MLE) of (α, β), denoted by $(\hat{\alpha}, \hat{\beta})$, can be attained by directly maximizing the likelihood function (2.1) or its log-transformation.

With the usual regularity conditions, $(\hat{\alpha}, \hat{\beta})$ are consistent and have asymptotic normality. That is, $\sqrt{n}(\hat{\alpha} - \alpha, \hat{\beta} - \alpha)'$ converges in distribution as $n \to \infty$ to the multivariate normal distribution with mean zero and variance $FI(\alpha, \beta)^{-1}$, where $FI(\alpha, \beta)$ is the Fisher information matrix. We estimate the asymptotic variance in our numerical studies using $-n^{-1}\partial^2 \log L(\alpha, \beta; N|T, Z)/\partial(\alpha, \beta)^2$, which converges to $FI(\alpha, \beta)$ almost surely as $n \to \infty$.

The computing needed to attain $(\hat{\alpha}, \hat{\beta})$ can be rather intensive, especially when extending to multiple latent classes. The following presents an application of the EM-algorithm (Dempster *et al.*, 1977), which is intuitive and easier to implement.

The log-likelihood function of (α, β) with the "full data" $\{(T_i, N_i, \eta_i, Z_i) : i = 1, \ldots, n\}$ is

$$l(\alpha, \beta; N, \eta|T, Z) = l_1(\alpha; \eta|Z) + l_2(\beta; N, \eta|T, Z)$$

with

$$l_1(\alpha; \eta|Z) = \sum_{i=1}^{n} \Big[\eta_i \log p(Z_i; \alpha) + (1 - \eta_i) \log [1 - p(Z_i; \alpha)]\Big] \tag{2.2}$$

and

$$l_2(\beta; N, \eta|T, Z) = \sum_{i=1}^{n} \Big[\eta_i \log P(N_i|\eta_i = 1, T_i, Z_i; \beta) + (1 - \eta_i) \log P(N_i|\eta_i = 0, T_i, Z_i)\Big]. \tag{2.3}$$

Note that $Q(\alpha, \beta; \alpha^*, \beta^*) = E\Big\{l(\alpha, \beta; N, \eta|T, Z)\big|T, N, Z; \alpha^*, \beta^*\Big\}$ is the sum of $Q_1(\alpha; \alpha^*, \beta^*) = E\Big\{l_1(\alpha; \eta|Z)\big|T, N, Z; \alpha^*, \beta^*\Big\}$ and $Q_2(\beta; \alpha^*, \beta^*) = E\Big\{l_2(\beta; N, \eta|T, Z)\big|T, N, Z; \alpha^*, \beta^*\Big\}$. Maximizing $Q(\alpha, \beta; \alpha^*, \beta^*)$ with respect to α, β is equivalent to separately maximizing $Q_1(\alpha; \alpha^*, \beta^*)$ and $Q_2(\beta; \alpha^*, \beta^*)$ with respect to α and β, respectively. Note that both $l_1(\alpha; \eta|Z)$ and $l_2(\beta; N, \eta|T, Z)$ are linear functions of η_i's, and thus $Q_1(\alpha; \alpha^*, \beta^*)$ and $Q_2(\beta; \alpha^*, \beta^*)$ are the corresponding linear functions of $E(\eta_i|T_i, N_i, Z_i; \alpha^*, \beta^*)$. This leads to an algorithm of iteratively alternating between an E-step and an M-step until convergence: the E-step estimates η_i's with their conditional expectations using the current estimates of (α, β), and the M-step maximizes separately (2.2) and (2.3) to attain the new estimates of (α, β) using the estimates of η_i's most recently updated by the E-step.

Specifically, starting with initial values $\alpha^{(0)}$ and $\beta^{(0)}$, at the kth iteration of the algorithm with $k \geq 1$ and the $(k-1)$th estimates $\alpha^{(k-1)}$ and $\beta^{(k-1)}$, the algorithm updates the estimates as follows.

E-Step. For $i = 1, \ldots, n$, calculate $\eta_i^{(k)} = E\{\eta_i | T_i, N_i, Z_i; \alpha^{(k-1)}, \beta^{(k-1)}\}$ as

$$\frac{P(N_i|\eta_i = 1, T_i, Z_i; \beta^{(k-1)})p(Z_i; \alpha^{(k-1)})}{P(N_i|\eta_i = 1, T_i, Z_i; \beta^{(k-1)})p(Z_i; \alpha^{(k-1)}) + P(N_i|\eta_i = 0, T_i, Z_i)[1 - p(Z_i; \alpha^{(k-1)})]}.$$

M-Step. Obtain $\alpha^{(k)}$ and $\beta^{(k)}$ by separately maximizing $l_1(\alpha; \eta^{(k)}|Z)$ and $l_2(\beta; N, \eta^{(k)}|T, Z)$, respectively, which is equivalent to solving the estimating equations with mild regularity conditions:

$$\frac{\partial l_1(\alpha; \eta^{(k)}|Z)}{\partial \alpha} = \sum_{i=1}^{n} [\eta_i^{(k)} - p(Z_i; \alpha)] \frac{\partial p(Z_i; \alpha)/\partial \alpha}{p(Z_i; \alpha)[1 - p(Z_i; \alpha)]} = 0 \tag{2.4}$$

and

$$\frac{\partial l_2(\beta; N, \eta^{(k)}|T, Z)}{\partial \beta} = \sum_{i=1}^{n} \eta_i^{(k)} [N_i - \Lambda_1(T_i, Z_i; \beta)] \frac{\partial \Lambda_1(T_i, Z_i; \beta)/\partial \beta}{\Lambda_1(T_i, Z_i; \beta)} = 0. \tag{2.5}$$

This algorithm with the ZIP model coincides the estimation procedure presented in Hall and Shen (2010). We may follow the discussion in Hall and Shen (2010) to provide a variation of the EM algorithm in the presence of outliers. We can verify the required conditions that ensure the resulting sequence $\{(\alpha^{(k)}, \beta^{(k)}) : k = 1, 2, \ldots\}$ of the EM-algorithm converges to the MLE $(\hat{\alpha}, \hat{\beta})$ derived from $L(\alpha, \beta; N|T, Z)$ in (2.1).

The MLE $(\hat{\alpha}, \hat{\beta})$ derived above is in fact the solution to the following estimating equations:

$$\sum_{i=1}^{n} [E(\eta_i|T_i, N_i, Z_i; \alpha, \beta) - p(Z_i; \alpha)] \frac{\partial p(Z_i; \alpha)/\partial \alpha}{p(Z_i; \alpha)[1 - p(Z_i; \alpha)]} = 0 \tag{2.6}$$

and

$$\sum_{i=1}^{n} E(\eta_i|T_i, N_i, Z_i; \alpha, \beta)[N_i - \Lambda_1(T_i, Z_i; \beta)] \frac{\partial \Lambda_1(T_i, Z_i; \beta)/\partial \beta}{\Lambda_1(T_i, Z_i; \beta)} = 0, \tag{2.7}$$

where $E(\eta_i|T_i, N_i, Z_i; \alpha, \beta)$ is

$$\frac{P(N_i|\eta_i = 1, T_i, Z_i; \beta)p(Z_i; \alpha)}{P(N_i|\eta_i = 1, T_i, Z_i; \beta)p(Z_i; \alpha) + P(N_i|\eta_i = 0, T_i, Z_i)[1 - p(Z_i; \alpha)]}. \tag{2.8}$$

Because the sequence $\{\eta^{(k)} : k = 1, 2, \ldots\}$ from the E-step converges to $E(\eta_i|T_i, N_i, Z_i; \alpha, \beta)$. Thus an alternative procedure for computing the MLE $(\hat{\alpha}, \hat{\beta})$ is to directly solve (2.6) and (2.7) jointly.

2.2 Simulation Study

We conducted a simulation study with the proposed mixture Poisson model to examine finite sample properties of the MLEs derived in *Section 2.1* in efficiency and robustness. Specifically, we evaluated three MLE sets obtained using datasets generated from five different model specifications. The three MLEs are (1) the MLE derived from the likelihood with the full-data $\{(T_i, N_i, \eta_i, Z_i) : i = 1, \ldots, n\}$, (2) the MLE directly attained from the likelihood with the observed-data $\{(T_i, N_i, Z_i) : i = 1, \ldots, n\}$, and (3) the MLE obtained by the EM-algorithm in *Section 2.1*. The first set of MLE in fact cannot be evaluated practically as it requires observation of the latent indicator η. It is used as a reference to study the performance of the other two MLEs. The five data settings and their underlying models are described in the following. The data generation and analysis were carried out using the *R Statistical Software* package.

2.2.1 Data Generation

Each simulated dataset had $n = 500$ independent subjects from two latent classes, say, the at-risk and not-at-risk groups. We considered two potential risk factors: the binary variable *sex* and the continuous variable age at baseline (*age*), which is the age of subject at the beginning of the study.

Risk factors, latent indicator η, and observation time length T were generated following the distributions below, respectively. For $i = 1, \ldots, n$,

- $sex_i \overset{iid}{\sim} Bin(1, 1/2)$ for the indicator of male
- $age_i \overset{iid}{\sim} Beta(0.7, 0.8)$, according to *age* trend in CAYACS data (McBride *et al.*, 2011; Ma, 2009) and standardized age values were used to have compatible coefficient with *sex*
- $\eta_i \overset{iid}{\sim} Bin(1, p_i)$, where $logit(p_i) = \alpha_0 + \alpha_1 sex_i + \alpha_2 age_i$, $\alpha_0 = 1$, $\alpha_1 = -1$ and $\alpha_2 = -0.8$
- $T_i \overset{iid}{\sim} Beta(2, 1) * 5$

We examine efficiency and robustness of the three MLEs in the following two simulation settings, respectively. Conditional on (η_i, T_i, Z_i), the observed number of physician visits associated with subject i was generated independently in two settings described as follows.

Simulation Setting 1:

This setting was designed to study efficiency of the estimation procedures. Conditional on (T_i, Z_i), the observed number of physician visits associated with subject i was generated independently from a Poisson distribution conditional on η_i, i.e., $N_i \overset{iid}{\sim} Poisson\Big(\Lambda_{\eta_i}(T_i, Z_i)\Big)$ given η_i. Two cases were simulated.

CASE 1A. The counts N_i's were from a ZIP model, with $N_i \equiv 0$ for subjects in the not-at-risk group (i.e., with $\eta_i = 0$) and in the at-risk group $N_i|\eta_i = 1 \overset{iid}{\sim} Poisson\Big(\Lambda_1(T_i, Z_i; \beta)\Big)$, $\Lambda_1(T_i, Z_i; \beta) = T_i^{\beta_3} \exp(\beta_0 + \beta_1 sex_i + \beta_2 age_i)$, where $\beta_0 = 1.8$, $\beta_1 = -0.6$, $\beta_2 = -0.5$ and $\beta_3 = 1$.

CASE 1B. The counts N_i's were from a mixture of two Poisson distributions for both classes, with $\Lambda_0(T_i, Z_i; \theta) = T_i^{\theta_3} \exp(\theta_0 + \theta_1 sex_i + \theta_2 age_i)$, $\theta_0 = 0.5$, $\theta_1 = -0.3$, $\theta_2 = -0.25$ and $\theta_3 = 1$ and $\Lambda_1(T_i, Z_i; \beta)$ as specified before.

Simulation Setting 2:

We designed the second setting to assess robustness of the estimating procedures. Conditional on (T_i, Z_i), the observed number of physician visits associated with subject i, N_i, in the not-at-risk group (i.e., $\eta_i = 0$) was generated from the Poisson distribution with mean $\Lambda_0(T_i, Z_i; \theta)$ as CASE 1B above. Three simulation cases were considered $N_i|\eta_i = 1, T_i, Z_i$ for subjects in the at-risk group generated from different distributions with the same mean function $\Lambda_1(T_i, Z_i; \beta)$ as *Setting 1* above.

CASE 2A. $N_i|\eta_i = 1 \overset{iid}{\sim} Bin(9, P_i)$ with $P_i = \Lambda_1(T_i, Z_i; \beta)/9$. This case simulates under-dispersed counts in the at-risk group with variance $\Lambda_1(T_i, Z_i; \beta)(1 - \Lambda_1(T_i, Z_i; \beta)/9)$.

CASE 2B. $N_i|\eta_i = 1 \overset{iid}{\sim} NB(size, P_i)$ with $P_i = size/(size + \Lambda_1(T_i, Z_i; \beta))$ and $size$ is the number of successful trials, or the dispersion parameter of a negative binomial. This case simulates over-dispersed counts in the at-risk group with variance $\Lambda_1(T_i, Z_i; \beta)(1 + \frac{\Lambda_1(T_i, Z_i; \beta)}{size})$. The over-dispersion gets worse as $size$ gets small. When $size$ goes to infinity, the negative binomial distribution converges to a Poisson distribution. We looked into cases with $size = 100$, 10, and 1.

CASE 2C. $N_i|\eta_i = 1 \overset{iid}{\sim} Poisson\Big(\xi_i \Lambda_1(T_i, Z_i; \beta)|\xi_i\Big)$. Here ξ_i is gamma-distributed with mean 1 and variance $1/\phi$. This case simulates over-dispersed counts in the at-risk group with variance $\Lambda_1(T_i, Z_i; \beta)(1 + \frac{1}{\phi})$. We used $\phi = 2$, 1 and 0.5 to simulate the counts from the at-risk group with over-dispersion levels ranging from low to high.

2.2.2 Simulation Results

We generated datasets in each of CASES 1A-1B and CASES 2A-2C, and evaluated the three sets of MLE for $(\alpha_0, \alpha_1, \alpha_2)$ and $(\beta_0, \beta_1, \beta_2, \beta_3)$, assuming the true values of θ are known. Based on 100 repetitions of the estimates, *Tables 2.1 - 2.4* present the sample means, the sample standard deviations, and the sample means of the asymptotic standard deviation estimates of the three estimators: the MLE by directly maximizing the observed-data likelihood (2.1) from $\{(N_i, T_i, Z_i) : i = 1, \ldots, n\}$ and the MLE via the EM algorithm described in *Section 2.1*, along with the MLE using "the full data", $\{(N_i, \eta_i, T_i, Z_i) : i = 1, \ldots, n\}$, which is not practically attainable. The third MLE is used as a reference to study the other two MLEs, which are theoretically the same.

Table 2.1 shows a summary of the simulation results with the data generated in CASES 1A-1B, following mixture Poisson distributions. All the sample means of the estimators are close to the corresponding true values of the parameters. This confirms that the three estimators are consistent. All the numerical outcomes with the MLE by directly maximizing the observed data likelihood are very close to the ones by the EM algorithm. The sample standard deviations associated with the MLEs using the observed data are slightly larger but quite close to the ones associated with the full data, which cannot be evaluated with the available data. We also verified the asymptotic normality for each of the three estimators, by examining the histograms of the attained estimates in the two simulation cases. *Table 2.1* includes also the sample means of the asymptotic standard deviation estimator for the MLEs in the two cases. The standard deviation estimates are close to the corresponding sample standard deviations of the MLEs. It indicates the standard deviation estimation is practically satisfactory with a reasonable sample size.

Tables 2.2 to *2.4* present the outcomes of the robustness study using the data generated from CASES 2A-2C. Those cases present certain model misspecification. The tables list MLEs and their stardard deviation estimates in the same way as *Table 2.1*. In general, the MLE with the full data shows certain robustness against model misspecification. The MLEs with the observed data, however, appear biased in the two cases with over-dispersion, and the bias gets heavier as over-dispersion gets worse. This indicates a great demand of a correct model specification especially when the membership of either groups is latent.

In CASE 2A with the under-dispersion data, the MLE with observed data seems acceptable, but the standard deviation estimators of all 3 sets of MLE are overestimated; see *Table*

2.2. We can see from the table that the sample means of the standard deviation estimates are much larger than the sample standard deviations of the estimates. On the other hand, in CASE 2B and CASE 2C, as over-dispersion is getting worse, not only are the MLEs with observed data rather biased, but also the standard deviation estimators of all 3 sets of MLE are underestimating. The sample means of standard deviation estimates are considerably smaller than the sample standard deviations. That is, the estimates differ a lot from sample to sample; see *Table 2.3 and 2.4*.

To graphically illustrate the lack of robustness of the likelihood-based estimation procedures against model misspecification of the at-risk group, we plot 3D and contour plots of the full-data log likelihood and the observed-data log likelihood surfaces for mixture Poisson data (CASE 1B) in *Figrue 2.1* and the same plots for Poisson($\eta = 0$) and mixed Poisson($\eta = 1$) data (CASE 2C), with $\phi = 1$ medium over-dispersion in *Figrue 2.2*. For demonstration purpose, data for these plots only include one covariate, *age*. Log likelihood functions are plotted against intercept α_0 and *age* coefficient α_1, when β's are fixed at true values. For data generated from mixture Poisson model (*Figure 2.1*), we can see clearly that both full-data likelihood and observed-data likelihood functions reach maximal points around the true values of (α_0, α_1). On the other hand, if the likelihood functions are misspecified, e.g., data generated from the Poisson and mixed Poisson model (CASE 2C), the full-data likelihood function still attains a maximal point in an acceptable range of (α_0, α_1); see *Figrue 2.2*. That is because η is observed and used to estimate α. However, the observed-data likelihood fails to attain a maximum in any reasonable range of (α_0, α_1).

Table 2.1: MLE of (α, β) in the Efficiency Study

	Case 1a: ZIP Model						
parameter	α_0	α_1(sex)	α_2(age)	β_0	β_1(sex)	β_2(age)	β_3(lnt)
true value	1	-1	-0.8	1.8	-0.6	-0.5	1
	MLE via full data						
sm[a]	1.009	-1.020	-0.794	1.794	-0.607	-0.504	1.006
ssd[b]	0.1793	0.1546	0.2972	0.0697	0.0378	0.0558	0.0488
sm^c_{sd}	0.1996	0.1891	0.3009	0.0731	0.0398	0.0572	0.0529
	MLE via observed data						
sm[a]	1.012	-1.018	-0.800	1.795	-0.607	-0.503	1.005
ssd[b]	0.1778	0.1515	0.2919	0.0702	0.0374	0.0556	0.0487
sm^c_{sd}	0.1996	0.1891	0.3009	0.0732	0.0398	0.0572	0.0529
	MLE via observed data by EM-algorithm						
sm[a]	1.013	-1.021	-0.801	1.795	-0.607	-0.504	1.005
ssd[b]	0.1827	0.1543	0.3000	0.0717	0.0375	0.0556	0.0498
sm^c_{sd}	0.1996	0.1892	0.3009	0.0732	0.0398	0.0572	0.0529
	Case 1b: Mixture Poisson						
parameter	α_0	α_1(sex)	α_2(age)	β_0	β_1(sex)	β_2(age)	β_3(lnt)
true value	1	-1	-0.8	1.8	-0.6	-0.5	1
	MLE via full data						
sm[a]	1.046	-1.010	-0.855	1.799	-0.597	-0.496	1.000
ssd[b]	0.2175	0.1786	0.2809	0.0808	0.0357	0.0565	0.0558
sm^c_{sd}	0.1996	0.1879	0.3002	0.0728	0.0395	0.0566	0.0522
	MLE via observed data						
sm[a]	1.042	-1.012	-0.821	1.797	-0.598	-0.501	1.003
ssd[b]	0.2353	0.2295	0.3285	0.0851	0.0493	0.0658	0.0597
sm^c_{sd}	0.2316	0.2495	0.3808	0.0837	0.0510	0.0656	0.0598
	MLE via observed data by EM-algorithm						
sm[a]	1.043	-1.0148	-0.820	1.797	-0.598	-0.500	1.003
ssd[b]	0.2353	0.2309	0.3231	0.0888	0.0489	0.0664	0.0628
sm^c_{sd}	0.2317	0.2494	0.3809	0.0838	0.0511	0.0657	0.0598

[a] Sample mean of the estimates

[b] Sample standard deviation of the estimates

[c] Sample mean of the standard deviation estimates

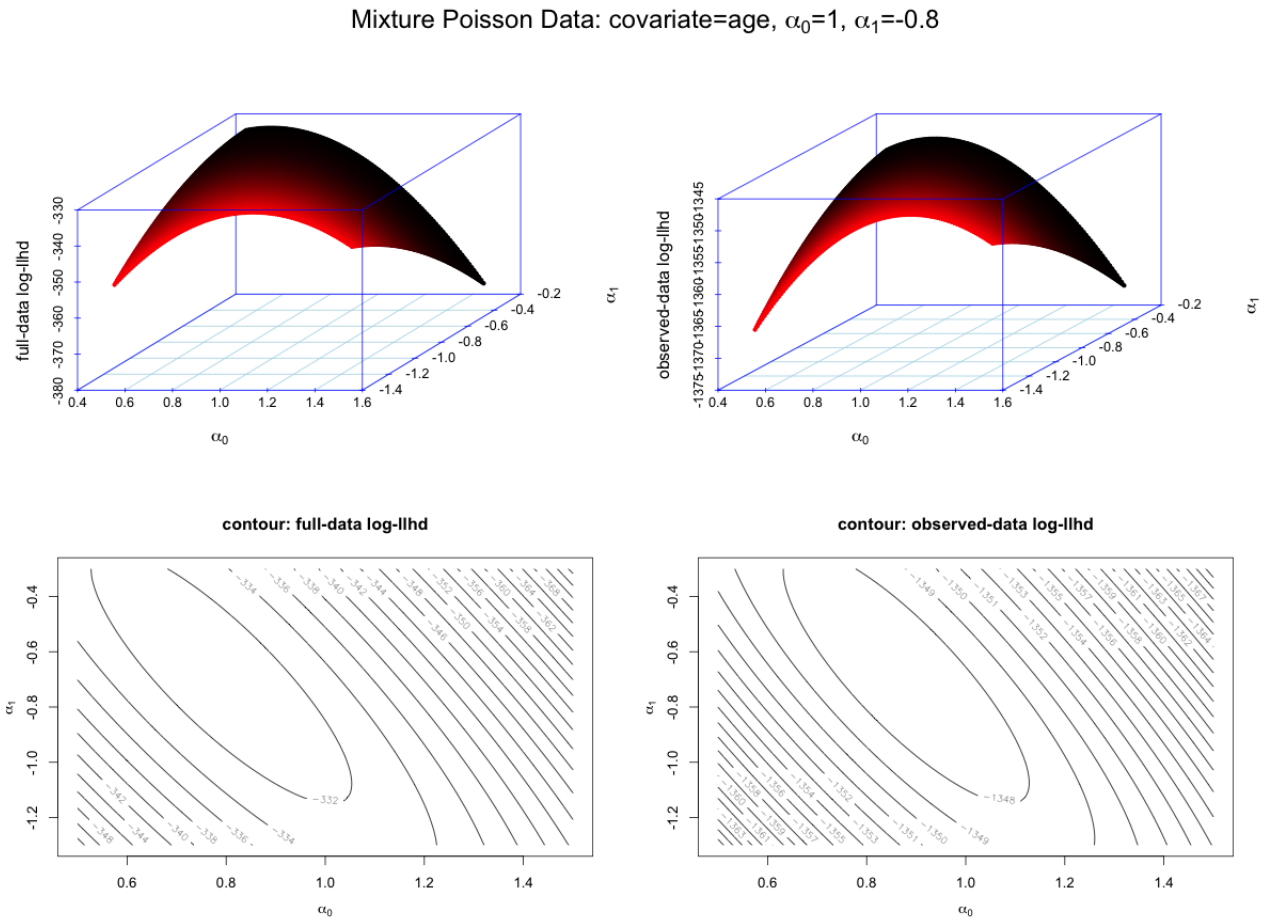

Figure 2.1: 3D and Contour Plots of Likelihood Surface for Case 1b

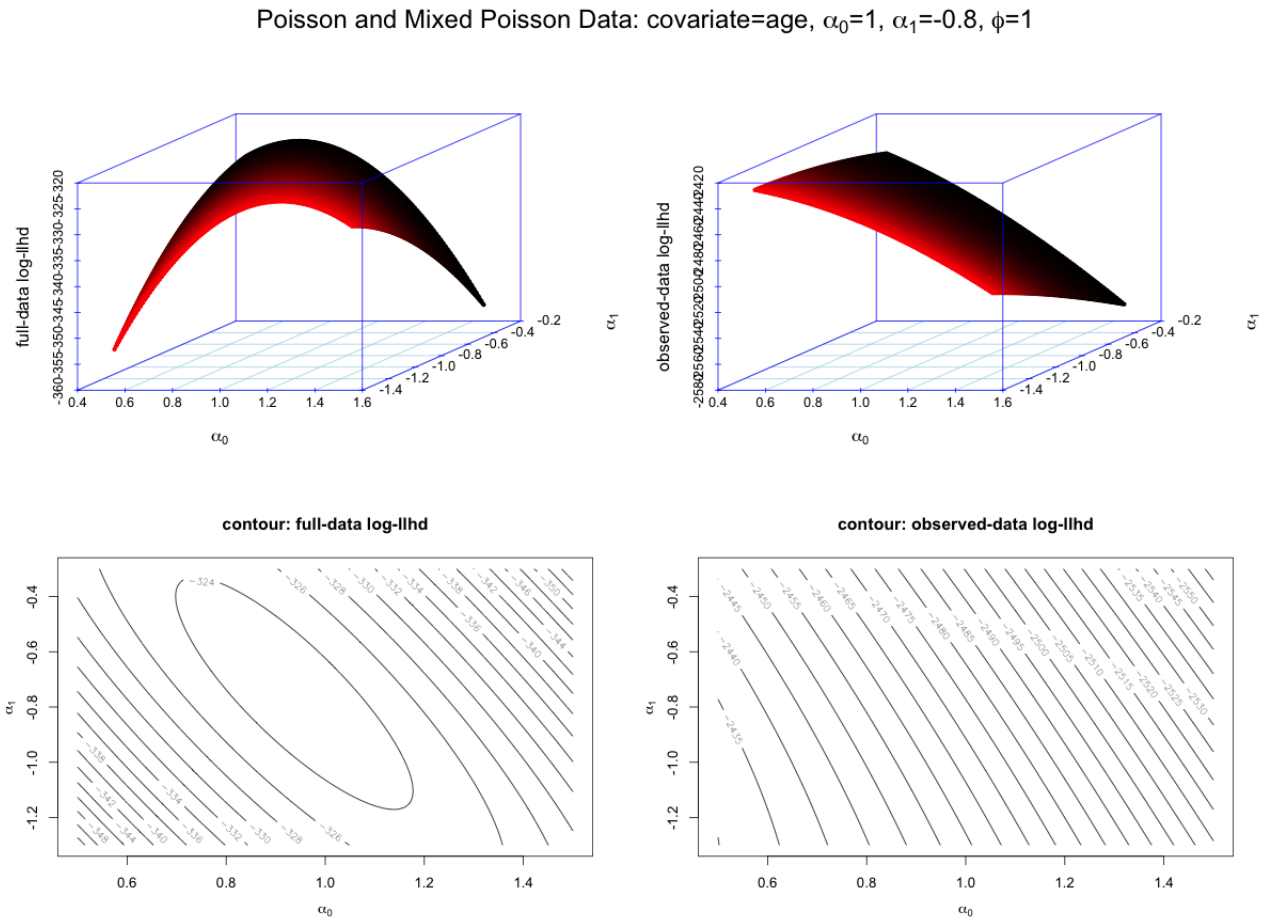

Figure 2.2: 3D and Contour Plots of Likelihood Surface for Case 2c

Table 2.2: MLE of (α, β) in the Robustness Study CASE 2A Poisson and Binomial

	CASE 2A: under-dispersion						
parameter	α_0	α_1(sex)	α_2(age)	β_0	β_1(sex)	β_2(age)	β_3(lnt)
true value	1	-1	-0.8	1.8	-0.6	-0.5	1
	MLE via full data						
sm[a]	1.006	-1.016	-0.798	1.848	-0.602	-0.506	0.977
ssd[b]	0.1949	0.1913	0.3134	0.0431	0.0315	0.0401	0.0321
sm^c_{sd}	0.2276	0.2410	0.3726	0.1097	0.0605	0.0860	0.0773
	MLE via observed data						
sm[a]	1.091	-0.931	-0.736	1.804	-0.634	-0.527	1.007
ssd[b]	0.2126	0.2390	0.3682	0.0467	0.0407	0.0473	0.0354
sm^c_{sd}	0.2321	0.2523	0.3856	0.1039	0.0582	0.0838	0.0728
	MLE via observed data by EM-algorithm						
sm[a]	1.090	-0.935	-0.728	1.804	-0.634	-0.528	1.008
ssd[b]	0.2154	0.2420	0.3718	0.0483	0.0404	0.0478	0.0361
sm^c_{sd}	0.2321	0.2523	0.3858	0.1038	0.0582	0.0837	0.0728

2.3 Discussion

We conclude from the simulation results in *Section 2.2.2* that the likelihood-based estimation procedures in this chapter are efficient with data from mixture Poisson distributions, but lack robustness against model misspecification. They may lead to biased estimates when the data have considerately heavy over-dispersion. Plus, in either of the under- or over-dispersion cases, the conventional standard deviation estimator can be biased.

This indicates a demand of robust inference procedures using counts with latent classes. Wang *et al.* (2012) presents a class of extended generalized estimating equations. The estimating equation based approach, which is equivalent to the likelihood-based approach when the counts are from the mixture Poisson model, requires only to specify the conditional mean function of $N_i|\eta_i = 1$, i.e., $\Lambda_1(T_i, Z_i; \beta)$, and the variance function.

In this chapter, we assume the true distribution of N in the not-at-risk group is fully known. This is conjectured from the assumption that the not-at-risk group has exactly the same physician visit pattern as the general population, and the fact that the general population has much richer information than the survivor cohort. In fact, at most there is a reasonable estimator of the distribution with the available information about the general population. It can result in undesirable inferences if one ignores the deviation of the

Table 2.3: MLE of (α, β) in the Robustness Study CASE 2B Poisson and Negative Binomial

parameter	α_0	α_1(sex)	α_2(age)	β_0	β_1(sex)	β_2(age)	β_3(lnt)
true value	1	-1	-0.8	1.8	-0.6	-0.5	1
	CASE 2B: $size = 100$						
	MLE via full data						
sm[a]	1.014	-1.010	-0.821	1.803	-0.601	-0.502	0.997
ssd[b]	0.2066	0.2042	0.2754	0.0729	0.0353	0.0631	0.0487
sm^c_{sd}	0.1995	0.1879	0.2475	0.0690	0.0341	0.0605	0.0481
	MLE via observed data						
sm[a]	0.989	-1.020	-0.826	1.806	-0.595	-0.499	0.998
ssd[b]	0.2334	0.2550	0.3997	0.0850	0.0462	0.0716	0.0583
sm^c_{sd}	0.2309	0.2484	0.3802	0.0811	0.0493	0.0637	0.0578
	MLE via observed data by EM-algorithm						
sm[a]	0.990	-1.022	-0.828	1.805	-0.595	-0.499	0.999
ssd[b]	0.2377	0.2582	0.4075	0.0874	0.0462	0.0714	0.0598
sm^c_{sd}	0.2309	0.2484	0.3803	0.0811	0.0494	0.0637	0.0578
	CASE 2B: $size = 10$						
	MLE via full data						
sm[a]	0.984	-0.991	-0.768	1.814	-0.607	-0.506	0.989
ssd[b]	0.1885	0.1825	0.3096	0.1066	0.0535	0.0941	0.0744
sm^c_{sd}	0.1863	0.1776	0.2991	0.0516	0.0267	0.0361	0.0365
	MLE via observed data						
sm[a]	0.632	-1.058	-0.820	1.884	-0.541	-0.445	0.977
ssd[b]	0.2435	0.2550	0.4275	0.1182	0.0680	0.1183	0.0811
sm^c_{sd}	0.2210	0.2366	0.3661	0.0699	0.0418	0.0491	0.0494
	MLE via observed data by EM-algorithm						
sm[a]	0.650	-1.063	-0.876	1.882	-0.541	-0.443	0.979
ssd[b]	0.2424	0.2590	0.4367	0.1197	0.0663	0.1182	0.0818
sm^c_{sd}	0.2212	0.2368	0.3666	0.0699	0.0418	0.0492	0.0494
	CASE 2B: $size = 1$						
	MLE via full data						
sm[a]	1.041	-1.028	-0.850	1.772	-0.612	-0.506	1.015
ssd[b]	0.1949	0.1843	0.2772	0.2378	0.1300	0.2382	0.1802
sm^c_{sd}	0.1975	0.1826	0.2568	0.0234	0.0112	0.0147	0.0166
	MLE via observed data						
sm[a]	-0.664	-0.960	-0.772	2.449	-0.448	-0.408	0.968
ssd[b]	0.3445	0.3409	0.5568	0.2734	0.1617	0.2992	0.2043
sm^c_{sd}	0.2297	0.2595	0.4014	0.0436	0.0227	0.0276	0.0309
	MLE via observed data by EM-algorithm						
sm[a]	-0.728	-0.954	-0.688	2.460	-0.449	-0.417	0.962
ssd[b]	0.2474	0.2895	0.4304	0.2746	0.1621	0.2940	0.2075
sm^c_{sd}	0.2309	0.2603	0.4029	0.0437	0.0228	0.0276	0.0310

Table 2.4: MLE of (α, β) in the Robustness Study CASE 2C Poisson and Mixed Poisson

parameter	α_0	α_1(sex)	α_2(age)	β_0	β_1(sex)	β_2(age)	β_3(lnt)
true value	1	-1	-0.8	1.8	-0.6	-0.5	1
	CASE 2C: $\phi = 2$						
	MLE via full data						
sm[a]	0.979	-0.991	-0.748	1.844	-0.606	-0.543	0.973
ssd[b]	0.1943	0.1850	0.2832	0.1636	0.1180	0.1674	0.1138
sm^c_{sd}	0.1796	0.1775	0.2895	0.0309	0.0152	0.0199	0.0219
	MLE via observed data						
sm[a]	-0.235	-0.976	-0.819	2.239	-0.474	-0.440	0.945
ssd[b]	0.2656	0.3000	0.4609	0.1901	0.1247	0.1774	0.1343
sm^c_{sd}	0.2150	0.2380	0.3675	0.0519	0.0292	0.0351	0.0369
	MLE via observed data by EM-algorithm						
sm[a]	-0.297	-0.957	-0.716	2.245	-0.474	-0.446	0.943
ssd[b]	0.2317	0.2819	0.4024	0.1908	0.1227	0.1752	0.1360
sm^c_{sd}	0.2155	0.2382	0.3681	0.0520	0.0292	0.0351	0.0370
	CASE 2C: $\phi = 1$						
	MLE via full data						
sm[a]	0.993	-0.994	-0.809	1.771	-0.607	-0.486	1.001
ssd[b]	0.2044	0.1746	0.3002	0.2468	0.1475	0.2650	0.1778
sm^c_{sd}	0.1989	0.1876	0.2804	0.0234	0.0111	0.0147	0.0164
	MLE via observed data						
sm[a]	-0.695	-0.842	-0.859	2.453	-0.461	-0.369	0.956
ssd[b]	0.3431	0.3551	0.6578	0.3525	0.1763	0.3208	0.2423
sm^c_{sd}	0.2313	0.2581	0.4030	0.0428	0.0233	0.0282	0.0307
	MLE via observed data by EM-algorithm						
sm[a]	-0.765	-0.879	-0.714	2.460	-0.466	-0.377	0.954
ssd[b]	0.2923	0.2648	0.4733	0.3377	0.1790	0.3169	0.2326
sm^c_{sd}	0.2326	0.2592	0.4046	0.0429	0.0233	0.0282	0.0307
	CASE 2C: $\phi = 0.5$						
	MLE via full data						
sm[a]	1.002	-0.979	-0.811	1.815	-0.577	-0.542	0.979
ssd[b]	0.1798	0.1905	0.2759	0.3375	0.1813	0.3147	0.2247
sm^c_{sd}	0.1805	0.1896	0.2791	0.0181	0.0085	0.0115	0.0129
	MLE via observed data						
sm[a]	-0.896	-0.942	-1.100	2.828	-0.412	-0.332	0.883
ssd[b]	0.4562	0.4608	0.8281	0.4119	0.2053	0.3760	0.2826
sm^c_{sd}	0.2490	0.2855	0.4443	0.0331	0.0163	0.0223	0.0236
	MLE via observed data by EM-algorithm						
sm[a]	-1.150	-0.807	-0.722	2.852	-0.420	-0.366	0.875
ssd[b]	0.3039	0.3277	0.4938	0.3990	0.2033	0.3727	0.2738
sm^c_{sd}	0.2540	0.2866	0.4490	0.0332	0.0163	0.0223	0.0236

estimator from the general population to the true distribution.

Now let us see numerically what happens if we use a value of θ different from its true value to make inferences about (α, β). We repeated the simulation of CASE 1B, but used incorrect θ value $(0.3, -0.1, -0.5, 0.8)$ instead of the true values. Estimates of the parameters of interest (α, β) were obtained and presented in *Table 2.5*. We can see that the MLEs of (α, β) via the observed-data give biased estimates of α and underestimate the standard deviations of $\hat{\beta}$. The MLE via the full-data remains consistent, because with the observation of η, the information of θ is not needed in making inferences on (α, β). This is not applicable in real data analysis. The simulation result in *Table 2.5* tells us that a untrue value of θ may lead to undesirable inferences about the parameters of interest. We need to account for it what available is an estimate rather than the true value. We will present the required conditions of θ estimator to make valid inferences on (α, β) in Chapter 3.

Table 2.5: MLE of (α, β) with Incorrect θ^*

	CASE 1B: Mixture Poisson						
parameter	α_0	α_1(sex)	α_2(age)	β_0	β_1(sex)	β_2(age)	β_3(lnt)
true value	1	-1	-0.8	1.8	-0.6	-0.5	1
	MLE via full data						
sm[a]	1.014	-1.005	-0.809	1.814	-0.598	-0.498	0.988
ssd[b]	0.1929	0.1813	0.3016	0.0711	0.0444	0.0520	0.0528
sm^c_{sd}	0.1996	0.1878	0.2997	0.0721	0.0419	0.0548	0.0545
	MLE via observed data						
sm[a]	1.054	-0.570	-0.088	1.809	-0.680	-0.615	0.989
ssd[b]	0.2429	0.2872	0.3648	0.0874	0.0605	0.0733	0.0619
sm^c_{sd}	0.2388	0.2524	0.3920	0.0701	0.0375	0.0512	0.0495
	MLE via observed data by EM-algorithm						
sm[a]	1.005	-0.562	0.014	1.813	-0.679	-0.621	0.988
ssd[b]	0.2396	0.2710	0.3788	0.0861	0.0584	0.0716	0.0620
sm^c_{sd}	0.2387	0.2526	0.3933	0.0702	0.0375	0.0511	0.0495

* true $\theta = (0.5, -0.3, -0.25, 1)$; incorrect $\theta = (0.3, -0.1, -0.5, 0.8)$

Chapter 3

Inference Procedures with Two Latent Classes

Chapter 2 assumes that the distribution of N_i in the "not-at-risk" group (i.e. $\eta_i = 0$) is fully known, and the parameters of interest (α, β) are estimated with the true values of θ. In reality, it is virtually impossible to know the true distribution. As seen in the discussion in *Section 2.3*, using misspecified values of θ to make inferences about (α, β) may result in estimation bias of α and underestimate the standard deviation of $\hat{\beta}$ by the estimation procedures via the observed data proposed in Chapter 2 under latent class modeling.

In attempt to address this issue and thus to make it feasible to apply our approach in practice, this chapter proposes two other likelihood-based approaches. Without knowing the true distribution of $N|\eta = 0$, the most straightforward way is to estimate (α, β, θ) simultaneously by the maximum likelihood estimation via the observed data under the mixture Poisson distribution with the two latent classes. This approach does not require extra information over the original data, which are referred to as "primary data" in our context. It is feasible in practical situations. However, lower efficiency is expected since the number of parameters increases and the information in use is the same. Moreover, sometimes estimating θ is not one of our primary interests, such as in the CAYACS program. One may not be willing to lose the efficiency in estimating the parameters (α, β). There is often information about one of the two latent classes from a different resource. Using the extra data appropriately can help us achieve more efficient estimator for the parameters of interest. We call the additional dataset as "supplementary data". The supplementary

data in CAYACS program we described previously are the available information about the general population from BC medical insurance database, while the primary data are the information of the survivor cohort from the CAYACS database. In other applications, such as the one considered in Hu and Lawless (1996), the supplementary data may not be so rich, such that the variation of the attained estimate of θ using the supplementary information is considerably larger. In either of the cases, we need to account for the variation of the estimated θ in inferences on α and β.

This chapter has two goals: (1) to establish a likelihood-based approach to make inferences for both latent classes with only the primary data, and (2) to propose a pseudo likelihood-based approach with supplementary data.

3.1 Maximum Likelihood Estimation with the Primary Data

3.1.1 Likelihood Functions and Estimating Procedures

Extending the model specification in *Section 2.1* for a mixture of two Poisson distributions, we further specify the mean function of $N|\eta = 0$ up to parameter θ, defined as $\Lambda_0(T, Z; \theta)$. Consider θ as another set of parameters needed to be estimated, and take the loglinear model formulation for the not-at-risk group: conditional on (T, Z), $log\Lambda_0(T, Z; \theta) = \theta_0 + \theta_1' Z + \theta_2 T$. The likelihood function of (α, β, θ) based on the observed data $\{(T_i, N_i, Z_i) : i = 1, \ldots, n\}$ is the same as the likelihood function (2.1) by plugging in the parametric form of $\Lambda_0(T, Z; \theta)$,

$$\begin{aligned} L(\alpha, \beta, \theta; N|T, Z) &= \prod_{i=1}^{n} \Big\{ P(N_i|\eta_i = 1, T_i, Z_i; \beta) P(\eta_i = 1|Z_i; \alpha) \\ &\quad + P(N_i|\eta_i = 0, T_i, Z_i; \theta)[1 - P(\eta_i = 1|Z_i; \alpha)] \Big\}, \end{aligned} \tag{3.1}$$

where $P(N_i|\eta_i = 0, T_i, Z_i; \theta) = \Lambda_0(T_i, Z_i; \theta)^{N_i} e^{-\Lambda_0(T_i, Z_i; \theta)} / N_i!$ and $P(N_i|\eta_i = 1, T_i, Z_i; \beta) = \Lambda_1(T_i, Z_i; \beta)^{N_i} e^{-\Lambda_1(T_i, Z_i; \beta)} / N_i!$. The MLE of (α, β, θ) may be attained by directly maximizing this likelihood function or its log-transformation.

An alternative way to attain the MLE is to extend the EM algorithm described in *Section 2.1*. The full-data log-likelihood functions (2.2) and (2.3) are updated as follows after including θ:

$$l(\alpha, \beta, \theta; N, \eta|T, Z) = l_1(\alpha; \eta|Z) + l_2(\beta; N, \eta|T, Z) + l_3(\theta; N, \eta|T, Z)$$

with

$$l_1(\alpha;\eta|Z)=\sum_{i=1}^{n}\Big[\eta_i\log p(Z_i;\alpha)+(1-\eta_i)\log\left[1-p(Z_i;\alpha)\right]\Big], \tag{3.2}$$

$$l_2(\beta;N,\eta|T,Z)=\sum_{i=1}^{n}\eta_i\log P(N_i|\eta_i=1,T_i,Z_i;\beta), \tag{3.3}$$

and

$$l_3(\theta;N,\eta|T,Z)=\sum_{i=1}^{n}(1-\eta_i)\log P(N_i|\eta_i=0,T_i,Z_i;\theta). \tag{3.4}$$

Similarly to the EM algorithm in *Section 2.1*, we iteratively alternate between the E-step, calculating $\eta_i^{(k)}=E\{\eta_i|T_i,N_i,Z_i;\alpha^{(k-1)},\beta^{(k-1)},\theta^{(k-1)}\}$ with the current (α,β,θ) estimate, and the M-step to maximize (3.2), (3.3), and (3.4) with the updated $\eta_i^{(k)}$ to update the current estimate of (α,β,θ) until convergence. Here, the computational advantage is quite obvious with θ, since the full data log-likelihood is the summation of (3.2), (3.3), and (3.4), each of which depends only on one set of the three sets of parameters.

3.1.2 Simulation Study

We conducted a simulation to examine finite samples of the above procedures. The simulation settings are as the same as in *Section 2.2.1*, except omitting Case 1a, which is the ZIP model and does not require estimation of θ. The two sets of MLE of (α,β,θ) via the observed data by directly maximizing (3.1) and the EM algorithm were evaluated, which are theoretically equivalent. As was done in Chapter 2, to obtain a comparison, we also evaluated the MLE of (α,β,θ) using the "full data", $\{(N_i,\eta_i,T_i,Z_i):i=1,\ldots,n\}$, which is not practically attainable. We estimated the asymptotic variance for the MLEs in the numerical study as the same as described in *Section 2.1*.

Tables 3.1 summarizes the simulation results based on 100 repetitions Case 1b, the mixture Poisson model, for the three MLEs of (α,β,θ) and their standard deviation estimates. The sample means are close to the corresponding true values of the parameters, which confirms that the three estimators are consistent. In *Table 2.1* of Chapter 2, the sample standard deviations associated with the MLE using the observed data are slightly larger but quite close to the ones associated with the ones using the full data. However, when one more set of parameters, θ, is estimated from the same simulation data setting, the standard deviations associated with the MLE using the observed data are much larger than the ones associated with the full data. This confirms the expectation of lower efficiency for

estimating all (α, β, θ) only from the primary data. In addition, the asymptotic standard deviation estimates of the MLEs were evaluated with the generated data. *Table 3.1* includes the sample means of the asymptotic standard deviation estimates for the MLEs, which are close to the corresponding sample standard deviations of the MLEs. This indicates the asymptotic standard deviations are practically satisfactory with a reasonable sample size.

For the robustness study Case 2a-2c in *Section 2.2.1*, we obtained similar results as in *Section 2.2.2* that the likelihood-based estimation procedures lack robustness against model misspecification and lead to biased estimates when the data have considerately heavy over-dispersion.

Table 3.1: MLE of (α, β, θ) in the Efficiency Study

	Case 1b: Mixture Poisson										
parameter	α_0	α_1(sex)	α_2(age)	β_0	β_1(sex)	β_2(age)	β_3(lnt)	θ_0	θ_1(sex)	θ_2(age)	θ_3(lnt)
true value	1	-1	-0.8	1.8	-0.6	-0.5	1	0.5	-0.3	-0.25	1
	MLE via full data										
sm[a]	1.007	-1.013	-0.804	1.796	-0.597	-0.503	1.003	0.494	-0.302	-0.256	1.004
ssd[b]	0.1908	0.1822	0.3071	0.0781	0.0395	0.0578	0.0542	0.1646	0.0731	0.1031	0.1012
sm^c_{sd}	0.1981	0.1876	0.2985	0.0710	0.0388	0.0555	0.0506	0.1485	0.0680	0.1067	0.1016
	MLE via observed data										
sm[a]	1.010	-1.005	-0.834	1.795	-0.596	-0.494	1.003	0.499	-0.308	-0.239	0.999
ssd[b]	0.2099	0.2533	0.3735	0.0757	0.0527	0.0670	0.0524	0.1788	0.0835	0.1163	0.1150
sm^c_{sd}	0.2418	0.2956	0.4172	0.0836	0.0566	0.0682	0.0593	0.1944	0.0923	0.1381	0.1310
	MLE via observed data by EM-algorithm										
sm[a]	1.003	-1.005	-0.837	1.793	-0.597	-0.497	1.005	0.502	-0.312	-0.253	1.003
ssd[b]	0.2315	0.2698	0.3970	0.0832	0.0533	0.0699	0.0584	0.2232	0.0918	0.1321	0.1376
sm^c_{sd}	0.2420	0.2959	0.4172	0.0838	0.0567	0.0683	0.0595	0.1945	0.0921	0.1381	0.1309

[a] Sample mean of the estimates

[b] Sample standard deviation of the estimates

[c] Sample mean of the standard deviation estimates

3.2 Pseudo Likelihood Estimation with Supplementary Data

Gong and Samaniego (1981) introduce the pseudo MLE for the parameter of interest, ρ, by plugging in an existing estimator of nuisance parameter, π, in the likelihood function of both ρ and π, and maximizing the resulted "pseudo-likelihood" function with respect to ρ. They also present conditions under which the pseudo MLE is consistent and asymptotically normal: (1) the pseudo MLE $\tilde{\rho}$ is consistent if the existing estimator $\hat{\pi}$ is; (2) the efficiency of $\tilde{\rho}$ depends on the relative efficiency of $\hat{\pi}$; (3) $\tilde{\rho}$ is asymptotically normal if $\hat{\pi}$ is $\sqrt{n}-$consistent and asymptotically normal. They investigate the numerical and asymptotic characteristics of the pseudo MLE under a particular functional form and derive asymptotic variances of the pseudo MLE. Hu and Lawless (1997) extend the pseudo MLE approach in Gong and Samaniego (1981) to situations where the nuisance parameter is an unknown distribution. They propose to use a nonparametric estimator for the unknown, such as the empirical distribution from a supplementary dataset. They also present an alternative to the pseudo MLE approach, an estimating function based approach in the situations.

As described from the beginning of this chapter, we assume that the not-at-risk group has the same physician visit pattern as the general population. By using information collected from the general population as the supplementary data, we can estimate the distribution of N_i in the not-at-risk group, and then follow the approaches in Gong and Samaniego (1981) and Hu and Lawless (1997), to verify the required conditions and derive the asymptotic properties of the pseudo MLE in our two latent classes modeling.

3.2.1 Estimating Procedure

As seen in *Section 2.3*, using an incorrect value of θ cannot achieve satisfactory inferences on (α, β). We need a consistent and asymptotically normal estimator of θ to make sure that the pseudo MLE of (α, β) has desirable asymptotic properties, such as consistency and asymptotic normality (Gong and Samaniego, 1981). Specifically, suppose that the sample size of the supplementary dataset used to estimate θ is M. With the usual regularity conditions, assume the estimator of θ, $\hat{\theta}$ is consistent and have asymptotic normality. That is, $\sqrt{M}(\hat{\theta} - \theta)$ converges in distribution as $M \to \infty$ to the multivariate normal distribution with mean zero and some variance function, $var(\theta)$. For example, if $\hat{\theta}$ is a MLE, $var(\theta)$ is equal to $FI(\theta)^{-1}$, where $FI(\theta)$ is the Fisher information matrix associated with the likelihood function of θ. If the supplementary data are heavily over-dispersed, $\hat{\theta}$ can be

the solution of quasi-Poisson score estimating equations, then $var(\theta)$ can be estimated by a sandwich estimator (e.g., Wang *et al.*, 2012).

By plugging in $\hat{\theta}$ to (3.1), we attain a pseudo-likelihood function of (α, β) based on the observed data $\{(T_i, N_i, Z_i) : i = 1, \ldots, n\}$, $L(\alpha, \beta, \hat{\theta}; N|T, Z)$. A pseudo MLE of (α, β) can be attained by directly maximizing this pseudo-likelihood function or its log-transformation with respect to (α, β). An alternative way is to extend the EM algorithm described in *Section 2.1*. The full-data pseudo log-likelihood function of (α, β) is obtained by plugging in $\hat{\theta}$ to the full-data log-likelihood function in *Section 3.1*:

$$l(\alpha, \beta, \hat{\theta}; N, \eta|T, Z) = l_1(\alpha; \eta|Z) + l_2(\beta; N, \eta|T, Z) + l_3(\hat{\theta}; N, \eta|T, Z),$$

where $l_1(\alpha; \eta|Z)$ and $l_2(\beta; N, \eta|T, Z)$ are as the same as in (3.2) and (3.3). The only differences from the EM algorithm described in *Section 3.1* are $\eta_i^{(k)} = E\{\eta_i|T_i, N_i, Z_i; \alpha^{(k-1)}, \beta^{(k-1)}, \hat{\theta}\}$ in the E-step and there is no need to maximize (3.4) in the M-step. These two sets of pseudo MLE are theoretically equivalent. Moreover, with the usual regularity conditions and a consistent and asymptotically normal estimator of θ, the pseudo MLE $(\tilde{\alpha}, \tilde{\beta})$ is consistent and asymptotically normal distributed. That is, $\sqrt{n}(\tilde{\alpha} - \alpha, \tilde{\beta} - \alpha)'$ converges in distribution as $n \to \infty$ and $M \to \infty$ to the multivariate normal distribution with mean zero and some variance function $V(\tilde{\alpha}, \tilde{\beta})$.

The important issue to be addressed for the pseudo MLE, different from the MLE proposed in Chapter 2 and *Section 3.1*, is estimating of the asymptotic variance of the pseudo MLE $(\tilde{\alpha}, \tilde{\beta})$, i.e., to estimate the variance function $V(\tilde{\alpha}, \tilde{\beta})$. We estimate θ and its asymptotic variance from the supplementary dataset, and get $\hat{\theta}$ and $\hat{var}(\hat{\theta})$ respectively. Obtained from plugging in $\hat{\theta}$ to the likelihood functions, the pseudo MLE of (α, β) are actually functions of $\hat{\theta}$, $\tilde{\alpha}(\hat{\theta})$ and $\tilde{\beta}(\hat{\theta})$. Theoretically, we can derive the variance function from the law of total variance,

$$V(\tilde{\alpha}, \tilde{\beta}) = E[V(\tilde{\alpha}, \tilde{\beta}|\hat{\theta})] + V[E(\tilde{\alpha}, \tilde{\beta}|\hat{\theta})]. \tag{3.5}$$

As the sample size of the supplementary dataset, $M \to \infty$, the second term of (3.5) converges to zero with some mild conditions, and the first term becomes the asymptotic variance of the MLE, which is $FI(\alpha, \beta)^{-1}$ derived in *Section 2.1*. Thus, theoretically, the asymptotic variance of the pseudo MLE is always larger than the one of the MLE. In our two latent classes modeling, if the supplementary dataset is fairly large, the variance estimate of the pseudo MLE will be close to the MLE, where assume the true θ is known as in Chapter

2. Otherwise, if the supplementary data are not rich, assuming the estimated θ is true may result in underestimating the variance of the estimator $(\tilde{\alpha}, \tilde{\beta})$.

3.2.2 Simulation Study

We conducted a simulation to study the pseudo MLE's finite sample properties. We generated data from Case 1b, the mixture Poisson distributions, in *Section 2.2.1*. The pseudo MLE of (α, β) via the observed data by directly maximizing and the EM algorithm were evaluated. This two sets of the pseudo MLE are theoretically equivalent. We assumed the existing estimator of θ to be consistent and asymptotically normal, i.e., $\hat{\theta} \sim N(\theta, sd^2)$. So $\hat{\theta}$ was generated from the normal distribution to estimate the pseudo MLE of (α, β). We also evaluated the MLE of (α, β) using the "full data", $\{(N_i, \eta_i, T_i, Z_i) : i = 1, \ldots, n\}$, which is not practically attainable. As seen before, the MLE of (α, β) and its variance estimation via full data does not need any information about θ, so it always has the asymptotic characteristics of MLE no matter what estimate of θ is used.

We used different values of sd in the distribution of $\hat{\theta}$ to investigate how the properties of the pseudo MLE of (α, β) and its variance estimation depends on the relative efficiency of $\hat{\theta}$. *Tables 3.2-3.3* summarize the simulation results based on 100 repetitions for the MLE via full data and two sets of the pseudo MLE of (α, β), as well as their standard deviation estimates. The pseudo MLEs for each dataset were evaluated under different realizations of $\hat{\theta}$ from its distribution. As a comparison, the standard deviation estimates of MLE for the pseudo MLEs were evaluated, assuming $\hat{\theta}$ as the true value. The standard deviation estimates of the pseudo MLEs (sd_{pMLE}) were obtained by bootstrap from B bootstrap samples according to (3.5). The first term of (3.5) was calculated from the mean of 100 bootstrap variances for each simulated dataset, and the second term of (3.5) was the variance of the 100 bootstrap sample means for each dataset. *Table 3.2* used sd equal to 30% of the magnitude of the true θ value and $B = 100$, and *Table 3.3* used 5% accordingly and $B = 200$.

In both tables, the true values of the parameters lie in the confidence intervals constructed by the pseudo MLE, which confirms the consistency of the pseudo MLE developed for two latent classes modeling, no matter the efficiency of $\hat{\theta}$. The results from both tables also confirm that the efficiency of the pseudo MLE is getting low with lower efficiency of $\hat{\theta}$ (Gong and Samaniego, 1981). We see from *Table 3.2*, when the sd of $\hat{\theta}$ is relatively large, if we assume $\hat{\theta}$ as true and take the variance estimation of the pseudo MLE as the same as MLE, we considerably underestimate the variation of parameter estimate. On the other

hand, when the sd of $\hat{\theta}$ is relatively small, the variance estimation of the pseudo MLE is so close to the MLE. This confirms our theoretical expectations and verifies the situation where the MLE presented in Chapter 2 is feasible.

Table 3.2: Pseudo MLE of (α, β) in the Efficiency Study
$\text{sd}(\hat{\theta}) = (0.15, 0.09, 0.075, 0.3)$; $B = 100$

	Case 1b: Mixture Poisson						
parameter	α_0	α_1(sex)	α_2(age)	β_0	β_1(sex)	β_2(age)	β_3(lnt)
true value	1	-1	-0.8	1.8	-0.6	-0.5	1
	MLE via full data						
sm[a]	0.998	-1.004	-0.781	1.787	-0.601	-0.505	1.012
ssd[b]	0.1799	0.1719	0.2983	0.0782	0.0377	0.0552	0.0519
sm^c_{sd}	0.1981	0.1873	0.2974	0.0720	0.0391	0.0557	0.0518
	pseudo MLE via observed data						
sm[a]	0.971	-0.866	-0.710	1.786	-0.668	-0.530	1.009
ssd[b]	0.3340	0.7259	0.5160	0.1215	0.2570	0.1253	0.0805
sm^c_{sd}	0.2410	0.2794	0.4091	0.0837	0.0560	0.0674	0.0599
sd^d_{pMLE}	0.4315	0.7875	0.6980	0.1566	0.2923	0.1392	0.1138
	pseudo MLE via observed data by EM-algorithm						
sm[a]	0.942	-0.862	-0.701	1.779	-0.669	-0.527	1.012
ssd[b]	0.3326	0.7528	0.6223	0.1221	0.2623	0.1252	0.0818
sm^c_{sd}	0.2418	0.2806	0.4113	0.0836	0.0564	0.0675	0.0599
sd^d_{pMLE}	0.4225	0.8308	0.7827	0.1549	0.2812	0.1491	0.1064

[a] Sample mean of the estimates
[b] Sample standard deviation of the estimates
[c] Sample mean of the standard deviation estimates of MLE
[d] Standard deviation estimates of the pseudo MLE

Table 3.3: Pseudo MLE of (α, β) in the Efficiency Study
$\text{sd}(\hat{\theta}) = (0.025, 0.015, 0.0125, 0.05)$; $B = 200$

	Case 1b: Mixture Poisson						
parameter	α_0	α_1(sex)	α_2(age)	β_0	β_1(sex)	β_2(age)	β_3(lnt)
true value	1	-1	-0.8	1.8	-0.6	-0.5	1
	MLE via full data						
sm[a]	1.028	-1.057	-0.796	1.791	-0.601	-0.494	1.006
ssd[b]	0.2187	0.1805	0.3195	0.0685	0.0401	0.0513	0.0482
sm^c_{sd}	0.1990	0.1882	0.2995	0.0720	0.0396	0.0566	0.0515
	pseudo MLE via observed data						
sm[a]	1.053	-1.062	-0.839	1.790	-0.601	-0.491	1.005
ssd[b]	0.2527	0.3036	0.3736	0.0758	0.0593	0.0683	0.0525
sm^c_{sd}	0.2323	0.2521	0.3819	0.0825	0.0516	0.0659	0.0591
sd^d_{pMLE}	0.3446	0.3997	0.5426	0.1130	0.0793	0.0950	0.0787
	pseudo MLE via observed data by EM-algorithm						
sm[a]	1.051	-1.063	-0.835	1.789	-0.601	-0.490	1.006
ssd[b]	0.2572	0.3040	0.3957	0.0789	0.0598	0.0693	0.0639
sm^c_{sd}	0.2324	0.2522	0.3820	0.0825	0.0516	0.0659	0.0591
sd^d_{pMLE}	0.3556	0.4068	0.5645	0.1157	0.0795	0.0955	0.0805

Chapter 4

Analysis of CAYACS Physician Visit Records

As aforementioned, this thesis project was motivated by a specific project of the CAYACS program, to evaluate the physician visit frequency of the survivor cohort, compare it with the general population, and identify risk factors. This chapter presents an analysis of the CAYACS physician visit data applying the proposed approaches.

4.1 Data Description and Preparation

The newly updated physician visit database includes physician visit records of 1962 individuals from the CAYACS cohort (Ma, 2009; McBride *et al.*, 2011), who were diagnosed before 20 years of age between January 1, 1981 and December 31, 1999, with a primary cancer or tumor, resided in British Columbia (BC) at the time of diagnosis, and have survived 5 or more years after diagnosis. We refer to this dataset as the "survivor cohort" dataset, which is our primary data. Since the BC Medical Services Plan (MSP) database started from January 1, 1986, the data collection period of each individual started from either this day or the day after 5 years diagnosis, whichever is later, and ended at either the day of individual death or December 31, 2006 , whichever is earlier. A randomly selected population sample of 19620 individuals was obtained from the client registry of the BC MSP database, known as the "general population" dataset and taken as the supplementary data in the following. People in the general population dataset who are at least 5 years of age,

registered with the provincial health insurance plan, and frequency-matched by birth year and sex to the survivor cohort (McBride *et al.*, 2011). Both datasets include physician visit records of general practice and specialist for each individual and their associated potential risk factors.

We chose to focus on general practice visits as the events of interest and analyzed the observed counts of the events. Potential risk factors were selected following the CAYACS's previous study (McBride *et al.*, 2011) but not including correlated factors, to avoid multicollinearity. For example, since type of cancer diagnosis and treatment of the cancer are correlated, and late-occurring health problems of survivors are often related to treatment (McBride *et al.*, 2011), we only included treatment in the analysis. Potential risk factors considered in the analysis are listed below:

- $S - sex$, the indicator of male.
- $A - age\ at\ baseline$, the age of individual at the beginning of the study and standardized values in the interval $(0, 1]$ were used. Note that $age\ at\ baseline = age\ at\ diagnosis + 5$ exactly for survivors.
- $SES - socioeconomic\ status$, the indicator of $ses = 4, 5(rich)$ according to Statistics Canada's census; $ses = 1, 2, 3(poor)$ otherwise.
- $C - treatment\ type$, the indicator of chemotherapy.
- $D - diagnosis\ period$, the indicator of diagnosis in $1990 - 1999(II)$; $1981 - 1990(I)$ otherwise.
- $T - time$, the length of individual's data collection period.

We applied the latent class model described in *Section 1.2*, where the model for the latent at-risk group indicator η is specified as

$$logit\Big(P(\eta = 1|S, A, SES, C, D)\Big) = \alpha_0 + \alpha_1 S + \alpha_2 A + \alpha_3 SES + \alpha_4 C + \alpha_5 D. \tag{4.1}$$

The frequency model for the at-risk group specifies the mean of the physician visit count conditional on $\eta = 1$ as

$$\begin{aligned} \Lambda_1(T, Z; \beta) &= E(N|\eta = 1, S, A, SES, C, D, T) \\ &= T^{\beta_6} \exp(\beta_0 + \beta_1 S + \beta_2 A + \beta_3 SES + \beta_4 C + \beta_5 D). \end{aligned} \tag{4.2}$$

Assuming the observed physician visit count in the not-at-risk group has the same distribution as in the general population, the frequency model for the not-at-risk group is:

$$\begin{aligned}\Lambda_0(T, Z; \theta) &= E(N|\eta = 0, S, A, SES, T) \\ &= T^{\theta_4} \exp(\theta_0 + \theta_1 S + \theta_2 A + \theta_3 SES).\end{aligned} \tag{4.3}$$

Here the risk factors C and D are not included because they are cancer survivor specific factors. We deleted individuals from both the survivor cohort and the general population datasets whose treatment type and/or ses are missing. The number of the individuals deleted is small, less than 2%.

As a preparation, we conducted a preliminary analysis of the counts by the conventional methods, Poisson regression and quasi Poisson regression, for both the general population and survivor cohort. The estimation results of parameters and their standard deviations from quasi Poisson regression for the general population are presented in the second and third columns of *Table 4.1*. This set of estimates served as a evaluation set of a consistent and asymptotically normal estimator of θ in (4.3), obtained from the supplementary data for estimating the pseudo MLE of α in the risk model (4.1) and β in the frequency model (4.2). Without noticing of the existence of two latent classes, the marginal mean of the physician visit counts in the survivor cohort can be specified as

$$\begin{aligned}\Lambda(T, Z; \beta^*) &= E(N|S, A, SES, C, D, T) \\ &= T^{\beta_6^*} \exp(\beta_0^* + \beta_1^* S + \beta_2^* A + \beta_3^* SES + \beta_4^* C + \beta_5^* D).\end{aligned}$$

Here, β^* has an interpretation different from β in (4.2). The estimates of β^* and their estimated standard deviations from quasi Poisson regression are presented in the last two columns of *Table 4.1*. We use this model and the associated parameter estimates to make a comparison with the analysis using the proposed approaches.

4.2 Analysis Results

We analyzed the real CAYACS data applying the proposed two latent classes model and the associated inference procedures. The analysis evaluated the portion of the survivors who are still at risk and thus have high frequency of physician visits, and identified the risk factors. We firstly estimated all the parameters in (4.1), (4.2) and (4.3) with only the survivor cohort dataset by the estimation procedure in *Section 3.1*. The MLE and its standard deviation

estimates were shown in the second and third columns of *Table 4.2*. By using the estimates of θ from the general population in *Table 4.1*, the pseudo MLE of (α, β) proposed in *Section 3.2* and its standard deviation estimator were evaluated. The estimates were presented in the last three columns of *Table 4.2*. The estimated standard deviation presented in the fifth column was calculated assuming the θ estimates as true values, and we call it as "sd_{naive}". The estimated standard deviation of the pseudo MLE (sd_{pMLE}) in the last column was attained by nonparametric bootstrap with 200 bootstrap samples, which took account for the variance estimation of $\hat{\theta}$ from the general population.

From *Table 4.1*, we see that both the general population and the survivor cohort data are heavily over-dispersed. All the risk factors considered for the number of general practice visits in the general population are statistically significant. In the quasi Poisson model for the survivor cohort, *sex*, *age at baseline* and *data collection length* are significantly associated with the marginal mean of the general practice visit counts.

The analysis of the survivor cohort data under the two latent classes model in *Table 4.2* provided more interesting results. The estimates of the regression parameters in the frequency model for counts in the not-at-risk group (see the MLE in *Table 4.2*), are quite close to the corresponding estimates from the general population in *Table 4.1*. This verified the CAYACS's previous finding and our assumption for the pseudo MLE procedure with the two latent classes model. All the potential risk factors in the frequency models for the at-risk group and the not-at-risk group are significant based on the MLE and its estimated standard deviations. But we also notice that, since the physician visit data in the survivor cohort are heavily over-dispersed, the standard deviation of MLE may be considerably underestimated. Plus, the parameter estimates may be biased, especially the estimated intercept terms, according to our simulation study for robustness in Chapter 2.

For the pseudo MLE in *Table 4.2*, the estimates of significant factors were in boldface according to the sd_{pMLE} column and itself was in boldface too. The sd_{naive} estimates were in boldface if the factors were significant according to them. Comparing the MLE and the pseudo MLE of (α, β) in *Table 4.2*, we see that they are quite consistent with each other, except the *sd* of MLE is significantly underestimated due to over-dispersion of the survivor cohort data, which results in more factors being significant. We also observe that the sd_{naive} considerably underestimates the standard deviation of parameter estimate, especially for β, which means the variance estimation of $\hat{\theta}$ from the general population is not negligible.

According to the pseudo MLE, $\tilde{\alpha}$, in the risk model and its estimated standard deviation,

we can evaluate the proportion of the at-risk group in the survivor cohort and identify its risk factors. For example, consider people at baseline in the risk model, who were female, age 5 at the beginning of the study, low *ses*, diagnosed cancer in the early stage *(1981-1990)*, and did not take chemotherapy. They have 19.4% to 36.0% chance to be in the at-risk group with a level of 95% confidence. The highest chance to be in the at-risk group happens to people who were female, age 25 at the beginning of the study, low *ses*, diagnosed cancer in the early stage, and took chemotherapy, which is 21.4% to 51.6% with a level of 95% confidence. The only significant factor for the risk model is the *diagnosis period*, which means that the survivors diagnosed in the early stage have significantly larger chance to be in the at-risk group. The pseudo MLE, $\tilde{\beta}$, in the frequency model and its estimated standard deviation can be used to identify significant factors associated with the high frequent physician visits in the at-risk group, and *sex* is the only risk factor detected by the pseudo MLE.

Most potential risk factors we considered in the risk model (*Table 4.2*) are not significant. Comparing to the interpretation of factors in the frequency model, risk factors in the risk model can be interpreted as follow. For example, the factor *sex* is significant in both the frequency models for the at-risk and the not-at-risk groups, but not significant in the risk model. This means that female has more frequent physician visits in general, no matter she is a cancer survivor or not. So *sex* is not a risk factor for the late and ongoing problems from cancer diagnosis. The two latent classes model leads to a natural comparison of the physician visits between the cancer survivors in the at-risk group and the general population.

Recall for the outcome of the robustness simulation study about the likelihood-based approaches in Chapter 2 under heavily over-dispersed data. Caution is required when further interpreting the data analysis. A robust inference procedure against model misspecification is in demand.

Table 4.1: Quasi Poisson Regression for the General Population and the Survivor Cohort[a]

	General Population		Survivor Cohort	
factor	*estimate*	*sd*	*estimate*	*sd*
the frequency model for counts				
intercept	**0.774**	(**0.0325**)	**1.398**	(**0.1379**)
male (vs female)	**-0.358**	(**0.0100**)	**-0.380**	(**0.0333**)
age at baseline	**0.396**	(**0.0265**)	**0.375**	(**0.0530**)
ses rich (vs poor)	**-0.048**	(**0.0104**)	-0.065	(0.0342)
chemo (vs not)			0.011	(0.0353)
diagnosis period II (vs I)			-0.007	(0.0489)
ln(time length)	**1.162**	(**0.0114**)	**0.961**	(**0.0464**)
dispersion parameter	15.10		14.52	

[a]Significant Effect with P-value ≤ 0.05 in **Boldface**

Table 4.2: Regression Parameters of Two Latent Classes Model for the Survivor Cohort[a]

	MLE		pseudo MLE		
factor	*estimate*	*sd*	*estimate*	sd_{naive}	sd^{b}_{pMLE}
the risk model for indicator $\eta = 1$					
intercept	-0.355	(0.1719)	**-1.001**	(**0.2045**)	(**0.2171**)
male (vs female)	-0.054	(0.1127)	-0.179	(0.1322)	(0.2171)
age at baseline	**0.656**	(**0.1778**)	0.334	(0.2135)	(0.3141)
ses rich (vs poor)	-0.004	(0.1122)	-0.107	(0.1316)	(0.1749)
chemo (vs not)	-0.072	(0.1191)	0.048	(0.1410)	(0.1528)
diagnosis period II (vs I)	**-0.617**	(**0.1125**)	**-0.470**	(**0.1339**)	(**0.2299**)
the frequency model for counts in the at-risk group					
intercept	**2.457**	(**0.0189**)	**2.480**	(**0.0229**)	(**0.2564**)
male (vs female)	**-0.291**	(**0.0045**)	**-0.193**	(**0.0064**)	(**0.0955**)
age at baseline	**0.200**	(**0.0078**)	0.213	(**0.0109**)	(0.2079)
ses rich (vs poor)	**-0.064**	(**0.0045**)	0.022	(**0.0064**)	(0.0621)
chemo (vs not)	**0.027**	(**0.0051**)	0.001	(0.0076)	(0.0672)
diagnosis period II (vs I)	**0.020**	(**0.0062**)	0.022	(**0.0082**)	(0.1050)
ln(time length)	**0.724**	(**0.0060**)	**0.742**	(**0.0072**)	(**0.0785**)
the frequency model for counts in the not-at-risk group			(from the general population)		
intercept	**0.780**	(**0.0207**)	**0.774**	(**0.0325**)	
male (vs female)	**-0.422**	(**0.0074**)	**-0.358**	(**0.0100**)	
age at baseline	**0.192**	(**0.0113**)	**0.396**	(**0.0265**)	
ses rich (vs poor)	**-0.056**	(**0.0076**)	**-0.048**	(**0.0104**)	
ln(time length)	**1.037**	(**0.0078**)	**1.162**	(**0.0114**)	

[a]Significant Effect with P-value ≤ 0.05 in **Boldface**

[b]Bootstrap Estimated Standard Deviation of the pseudo MLE

Chapter 5

Final Remarks

5.1 Summary

Motivated by the CAYACS program, this MSc thesis project proposes a latent class model to formulate count data from a cohort with potential two strata. In the young cancer survivor cohort, these two classes are the "at-risk" group who still suffer the consequence of cancer diagnosis and visit physicians more frequently and the "not-at-risk" group who have the same physician visit frequencies as the general population. We are interested in evaluating the proportion of the at-risk group, assessing the frequency of physician visits associated with the at-risk group, identifying the associated risk factors, and making comparison to the general population.

Under a mixture Poisson model, we present several likelihood-based inference procedures with or without one class fully specified. Finite sample efficiency and robustness properties of these procedures have been studied via simulation under different scenarios. We begin with assuming the distribution of the not-at-risk group as known and develop the maximum likelihood estimator of the model parameters for the at-risk group. The simulation results show that likelihood-based estimating procedures are quite efficient under the mixture Poisson model, but lack of robustness against model misspecification. Without fully specifying the not-at-risk group's distribution, we propose approaches to making inferences on both latent classes. The likelihood based approach is rather computationally intensive as expected. Thus, we propose a pseudo likelihood inference procedure by estimating the distribution of the not-at-risk group from the general population as well as taking account for the variance estimation. The required conditions are given for an estimator of nuisance

parameter from supplementary data to make inference on the parameters of interest. To illustrate our approaches, the real CAYACS data are analyzed. We identify the risk factors by the pseudo MLE for the risk and frequency models. Our approach provides a natural comparison of the at-risk group to the general population.

5.2 Future Investigation

Some finial comments and interesting points for future investigation are listed below.

Theoretically, directly maximizing the observed data likelihood and the EM algorithm are equivalent. Both simulation and the real data analysis showed the EM algorithm is more robust to the choice of initial values used in the numerical procedures. It could be interesting to further explore this.

Under the mixture Poisson model, (α, β, θ) always can be estimated theoretically. However, numerically, the group classification may be not easy if β and θ are too close. And in this situation, the two latent classes model may be unnecessary, since it almost triples the number of parameters. It could be interesting to theoretically derive or numerically study the range of parameters to make a latent class model appropriate.

The parametric form we specified for the means of the latent variable and the counts of both classes in *Section 1.2* is just a special case. For example, in Chapter 4, we showed *age at baseline* is not a significant risk factor in the frequency model for the at-risk group. But this result could be due to the misspecification of the systematic part of the frequency model. Quadratic or nonparametric form should be considered in the future for the right hand side of $log\{\Lambda_1(t, Z; \beta)\} = \beta_0 + \beta_1' Z + \beta_2 t$ for factor *age*.

We are aware from the simulation study in *Section 3.2* that the estimated standard deviation of the pseudo MLE from bootstrap seems systematically larger than the *ssd*. More study about the variance estimation of pseudo MLE for latent class models may be needed.

We can extend the approaches developed in this project in multiple aspects. For example, it is possible to extend the approaches to more than two/finite latent classes. An application could be that more frequent physician visits occur in a third group of cancer survivors that may be caused by fear. The count data we considered in this project are cross-sectional. This also can be easily extended to analyse longitudinal counts. According to one of the CAYACS's objectives, one could try to formulate physician visit cost data by latent class

models and develop corresponding inference procedures.

An interesting and important issue noticed from this project is that the likelihood-based estimating procedures for latent class models have unsatisfactory robustness against model misspecification. The simulation results indicate a strong demand of developing robust estimation procedures. It is frequently of demand to develop robust inferential procedures, particularly in epidemiological and medical applications. Wang *et al.* (2012) develop a robust estimating procedure for two latent class models without specifying the underlying distribution but only the mean and variance functions. The EM-algorithm discussed in *Section 2.1*, more specifically equations (2.6) and (2.7), motivated the extended GEE approach with estimating (2.8) non-parametrically. The approach is readily extendable to accommodate situations with clustered data.

Bibliography

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–38.

Gong, G. and Samaniego, F. (1981). Pseudo maximum likelihood estimation: theory and applications. *The Annals of Statistics*, pages 861–869.

Goodman, L. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, **61**(2), 215–231.

Hall, D. and Shen, J. (2010). Robust estimation for zero-inflated poisson regression. *Scandinavian Journal of Statistics*, **37**(2), 237–252.

Hu, X. and Lawless, J. (1996). Estimation from truncated lifetime data with supplementary information on covariates and censoring times. *Biometrika*, **83**(4), 747–761.

Hu, X. and Lawless, J. (1997). Pseudolikelihood estimation in a class of problems with response-related missing covariates. *Canadian Journal of Statistics*, **25**(2), 125–142.

Lambert, D. (1992). Zero-inflated poisson regression, with an application to defects in manufacturing. *Technometrics*, pages 1–14.

Lazarsfeld, P. and Henry, N. (1968). *Latent structure analysis*. Houghton, Mifflin.

Ma, S. (2009). Analyses of physician visits from childhood and adolescent cancer survivors.

Magidson, J. and Vermunt, J. (2002). Latent class models for clustering: A comparison with k-means. *Canadian Journal of Marketing Research*, **20**(1), 36–43.

McBride, M., Rogers, P., Sheps, S., Glickman, V., Broemeling, A., Goddard, K., Hu, J., Lorenzi, M., Peacock, S., Pritchard, S., *et al.* (2010). Childhood, adolescent, and young adult cancer survivors research program of british columbia: Objectives, study design, and cohort characteristics. *Pediatric blood & cancer*, **55**(2), 324–330.

McBride, M., Lorenzi, M., Page, J., Broemeling, A., Spinelli, J., Goddard, K., Pritchard, S., Rogers, P., and Sheps, S. (2011). Patterns of physician follow-up among young cancer survivors. *Canadian Family Physician*, **57**(12), e482–e490.

Pepe, M. and Janes, H. (2007). Insights into latent class analysis of diagnostic test performance. *Biostatistics*, **8**(2), 474–484.

Vermunt, J. (2008). Latent class and finite mixture models for multilevel data sets. *Statistical Methods in Medical Research*, **17**(1), 33–51.

Wang, H., Hu, X., Lorenzib, M., McBride, M., and Spinelli, J. (2012). Analysis of counts with two latent classes, with application to risk assessment using records of physician visit. *(submitted for publication)*.